

Multiple Regression Example

Applied Regression and Other Multivariable Methods
Sections 8-1 - 8-6

9

SAS Program

```
options nocenter;
goptions reset=global colors=(none);

data problem81;
infile 'i:\www\datasets502\EX0502.DAT' firstobs=2 dlm='09'x;
input person sbp quet age smk;

/* Compute frequency table for smk */
proc freq;
tables smk;
run;

/* Generate summary statistics and histograms for sbp, age, quet */
proc univariate;
var sbp quet age;
histogram sbp quet age / kernel;
run;

/* Generate scatterplots for several variable combinations*/
symbol1 v=circle i=none;
proc gplot;
plot sbp*quet sbp*age quet*age;
run;

/* Estimate regression lines */
proc reg;
model sbp quet = age;
model sbp = quet;
model sbp = smk;
run;
```

9-2

The Problem

Consider the hypothetical sample of 32 white males over the age of 40 from the town of Angina that was described in Chapter 5 (Prob #2). Suppose we're interested in describing the relationship between systolic blood pressure (sbp) and the following three independent variables

1. Body size (quet) : Defined to be $100(\text{weight}/\text{height}^2)$
2. Age (age) : Consider only individuals over the age of 40
3. Smoking history (smk) : If ever a smoker, $\text{smk}=1$

To address this problem, we'll do the following steps

1. Investigate each variable individually
2. Investigate pairwise relationships
3. Investigate different choices of models

9-1

```
/* Generate side by side boxplots (smk=0 and 1) for sbp, age, and quet */
proc sort; by smk;
proc boxplot;
plot (sbp quet age)*smk / pctldef=4;
run;

/* Generate scatterplots with diff regression lines for smk=0 and 1 */
symbol1 v=circle i=r1;
symbol2 v=plus i=r1 line=2;
proc gplot;
plot sbp*quet=smk sbp*age=smk;
run;

/* Create new independent variables to investigate */
data problem81;
set problem81;
quet_smk = quet*smk;
quet_age = quet*age;

/* Investigate different models */
proc reg;
model sbp = quet age smk quet_smk quet_age;
plot r.*p. /nostat legend=none;
plot r.*nqq. / nostat;
run;
delete quet_smk;
print;
run;
delete quet_age;
print;
run;
delete quet;
print;
run;
```

9-3

The Output

*** Over 50% of the sample smokes ***

The FREQ Procedure

| smk | Frequency | Percent | Cumulative Frequency | Cumulative Percent |
|-----|-----------|---------|----------------------|--------------------|
| 0 | 15 | 46.88 | 15 | 46.88 |
| 1 | 17 | 53.13 | 32 | 100.00 |

*** Summary statistics for SBP ***

The UNIVARIATE Procedure

Variable: sbp

| Moments | | | |
|-----------------|------------|------------------|------------|
| N | 32 | Sum Weights | 32 |
| Mean | 144.53125 | Sum Observations | 4625 |
| Std Deviation | 14.3975454 | Variance | 207.289315 |
| Skewness | 0.52712743 | Kurtosis | -0.1088519 |
| Uncorrected SS | 674883 | Corrected SS | 6425.96875 |
| Coeff Variation | 9.96154495 | Std Error Mean | 2.5451505 |

Basic Statistical Measures

| Location | | Variability | |
|----------|----------|---------------------|-----------|
| Mean | 144.5313 | Std Deviation | 14.39755 |
| Median | 143.0000 | Variance | 207.28931 |
| Mode | 132.0000 | Range | 60.00000 |
| | | Interquartile Range | 17.50000 |

9-4

*** Summary statistics for QUET and AGE ***

Variable: quet

| Moments | | | |
|-----------------|------------|------------------|------------|
| N | 32 | Sum Weights | 32 |
| Mean | 3.44109375 | Sum Observations | 110.115 |
| Std Deviation | 0.49707807 | Variance | 0.2470866 |
| Skewness | 0.17019173 | Kurtosis | -0.1861304 |
| Uncorrected SS | 386.575723 | Corrected SS | 7.65968472 |
| Coeff Variation | 14.4453509 | Std Error Mean | 0.08787182 |

Basic Statistical Measures

| Location | | Variability | |
|----------|----------|---------------------|---------|
| Mean | 3.441094 | Std Deviation | 0.49708 |
| Median | 3.380500 | Variance | 0.24709 |
| Mode | . | Range | 2.26900 |
| | | Interquartile Range | 0.76350 |

Variable: age

| Moments | | | |
|-----------------|------------|------------------|------------|
| N | 32 | Sum Weights | 32 |
| Mean | 53.25 | Sum Observations | 1704 |
| Std Deviation | 6.95608344 | Variance | 48.3870968 |
| Skewness | 0.08755885 | Kurtosis | -1.0224424 |
| Uncorrected SS | 92238 | Corrected SS | 1500 |
| Coeff Variation | 13.0630675 | Std Error Mean | 1.22967344 |

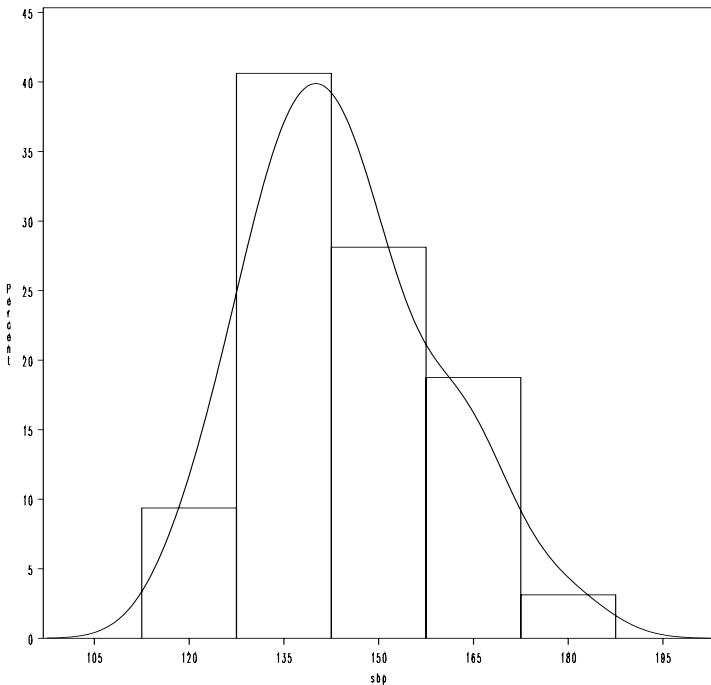
Basic Statistical Measures

| Location | | Variability | |
|----------|----------|---------------------|----------|
| Mean | 53.25000 | Std Deviation | 6.95608 |
| Median | 53.50000 | Variance | 48.38710 |
| Mode | 54.00000 | Range | 24.00000 |
| | | Interquartile Range | 10.50000 |

NOTE: The mode displayed is the smallest of 2 modes with a count of 3.

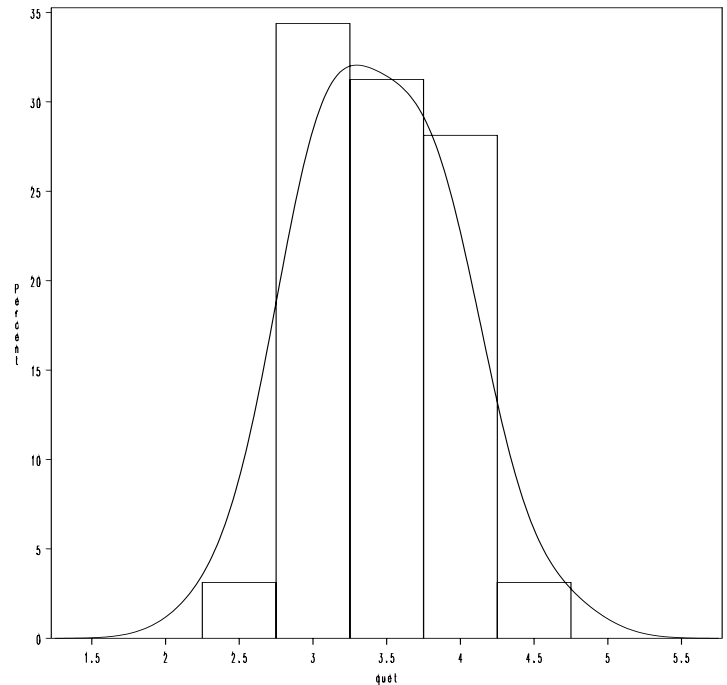
9-5

Histogram of SBP



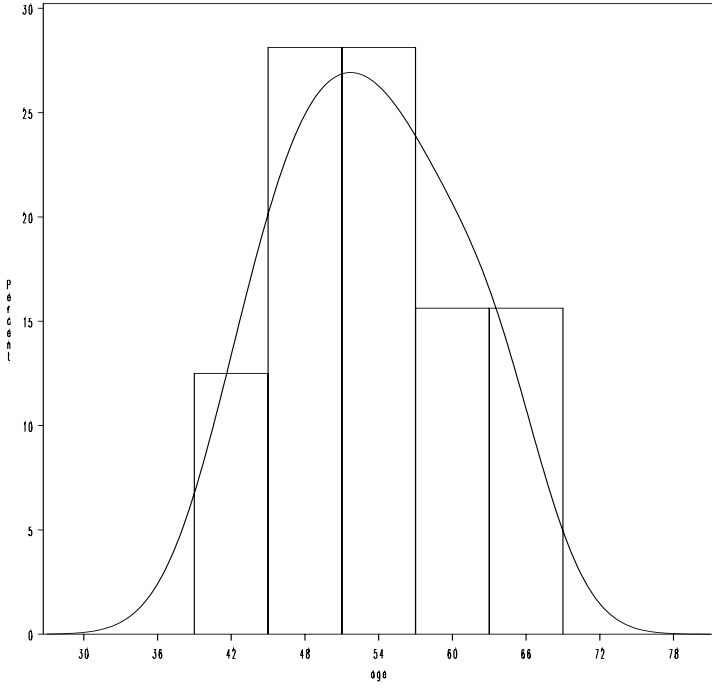
9-6

Histogram of QUET



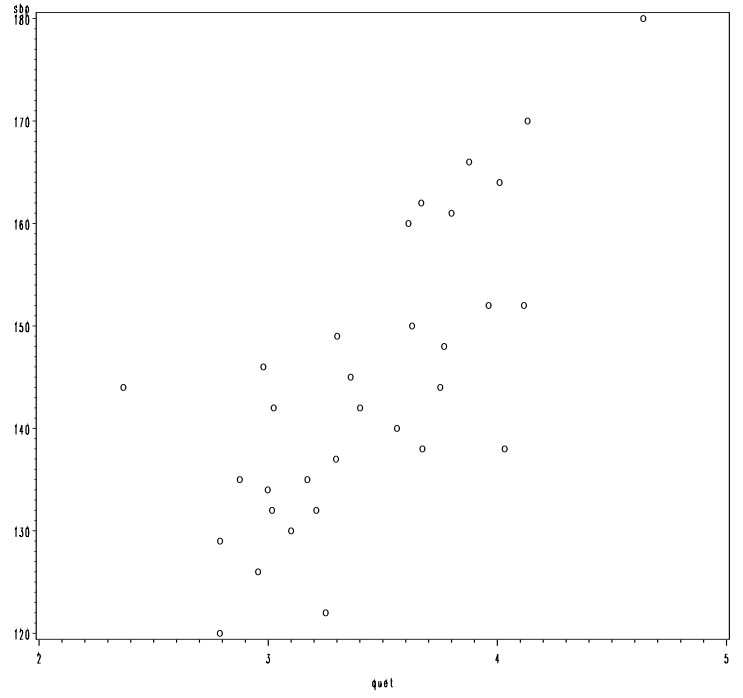
9-7

Histogram of AGE



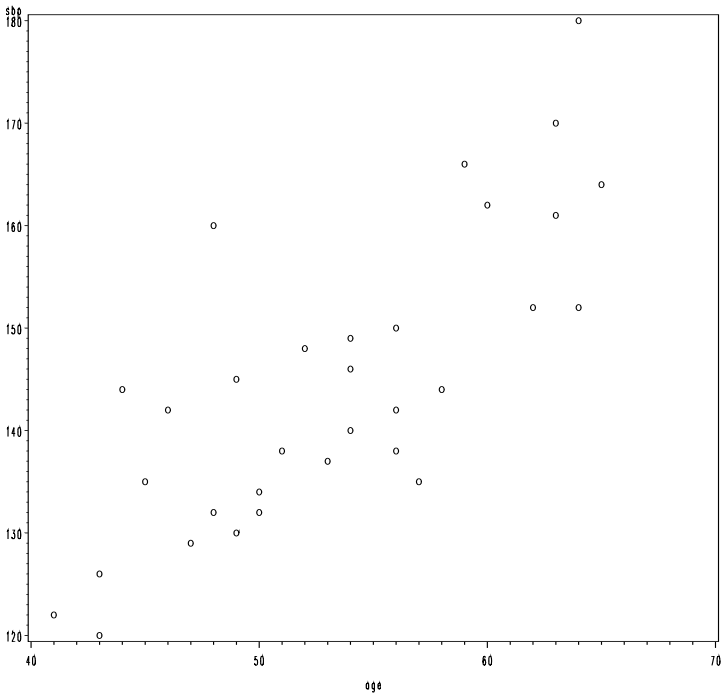
9-8

Scatterplot of SBP vs QUET



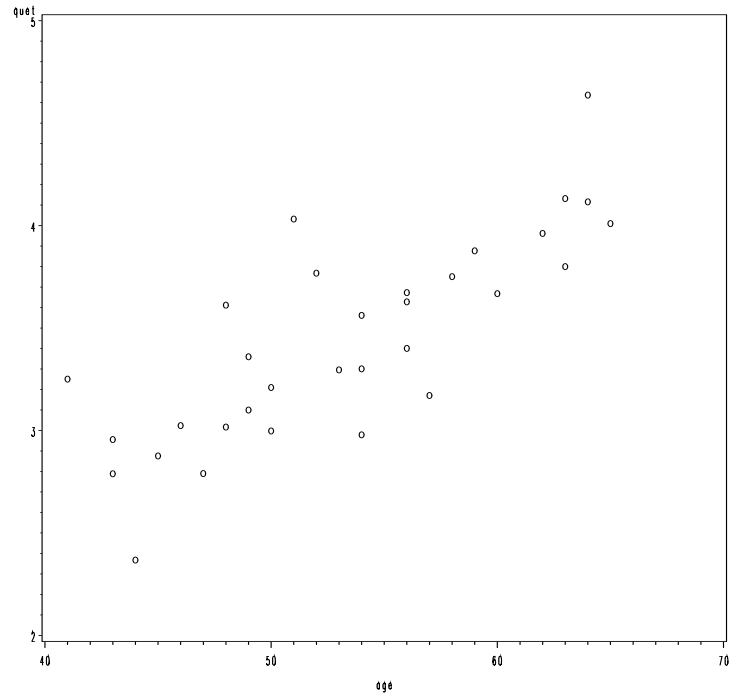
9-9

Scatterplot of SBP vs AGE



9-10

Scatterplot of QUET vs AGE



9-11

The REG Procedure

**** Dependent Variable: sbp

| Analysis of Variance | | | | | |
|----------------------|----|----------------|-------------|---------|--------|
| Source | DF | Sum of Squares | Mean Square | F Value | Pr > F |
| Model | 1 | 3861.63038 | 3861.63038 | 45.18 | <.0001 |
| Error | 30 | 2564.33838 | 85.47795 | | |
| Corrected Total | 31 | 6425.96875 | | | |

| Root MSE | 9.24543 | R-Square | 0.6009 |
|----------------|-----------|----------|--------|
| Dependent Mean | 144.53125 | Adj R-Sq | 0.5876 |
| Coeff Var | 6.39684 | | |

| Parameter Estimates | | | | | |
|---------------------|----|--------------------|----------------|---------|---------|
| Variable | DF | Parameter Estimate | Standard Error | t Value | Pr > t |
| Intercept | 1 | 59.09163 | 12.81626 | 4.61 | <.0001 |
| age | 1 | 1.60450 | 0.23872 | 6.72 | <.0001 |

**** Dependent Variable: quet

| Analysis of Variance | | | | | |
|----------------------|----|----------------|-------------|---------|--------|
| Source | DF | Sum of Squares | Mean Square | F Value | Pr > F |
| Model | 1 | 4.93597 | 4.93597 | 54.37 | <.0001 |
| Error | 30 | 2.72371 | 0.09079 | | |
| Corrected Total | 31 | 7.65968 | | | |

| Root MSE | 0.30131 | R-Square | 0.6444 |
|----------------|---------|----------|--------|
| Dependent Mean | 3.44109 | Adj R-Sq | 0.6326 |
| Coeff Var | 8.75636 | | |

| Parameter Estimates | | | | | |
|---------------------|----|--------------------|----------------|---------|---------|
| Variable | DF | Parameter Estimate | Standard Error | t Value | Pr > t |
| Intercept | 1 | 0.38645 | 0.41769 | 0.93 | 0.3622 |
| age | 1 | 0.05736 | 0.00778 | 7.37 | <.0001 |

9-12

**** Dependent Variable: sbp

| Analysis of Variance | | | | | |
|----------------------|----|----------------|-------------|---------|--------|
| Source | DF | Sum of Squares | Mean Square | F Value | Pr > F |
| Model | 1 | 3537.94574 | 3537.94574 | 36.75 | <.0001 |
| Error | 30 | 2888.02301 | 96.26743 | | |
| Corrected Total | 31 | 6425.96875 | | | |

| Root MSE | 9.81160 | R-Square | 0.5506 |
|----------------|-----------|----------|--------|
| Dependent Mean | 144.53125 | Adj R-Sq | 0.5356 |
| Coeff Var | 6.78856 | | |

| Parameter Estimates | | | | | |
|---------------------|----|--------------------|----------------|---------|---------|
| Variable | DF | Parameter Estimate | Standard Error | t Value | Pr > t |
| Intercept | 1 | 70.57640 | 12.32187 | 5.73 | <.0001 |
| quet | 1 | 21.49167 | 3.54515 | 6.06 | <.0001 |

**** Dependent Variable: sbp

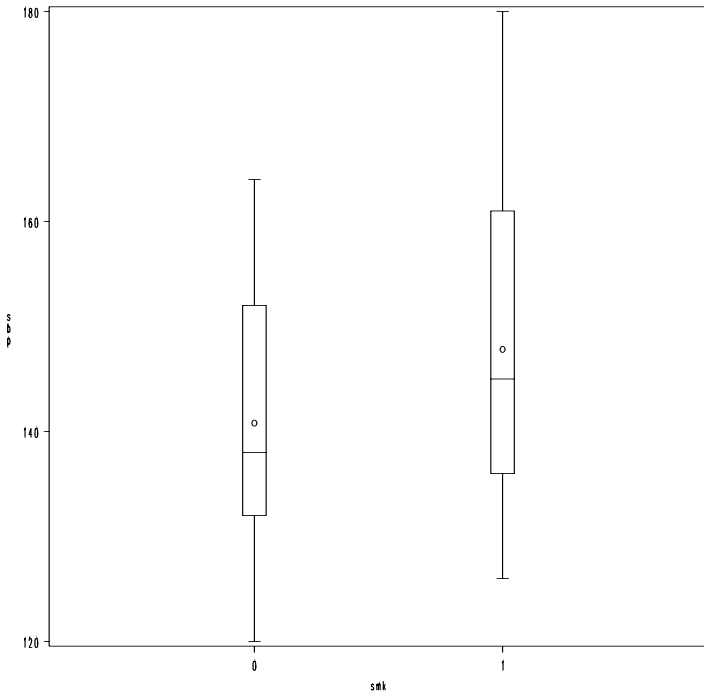
| Analysis of Variance | | | | | |
|----------------------|----|----------------|-------------|---------|--------|
| Source | DF | Sum of Squares | Mean Square | F Value | Pr > F |
| Model | 1 | 393.09816 | 393.09816 | 1.95 | 0.1723 |
| Error | 30 | 6032.87059 | 201.09569 | | |
| Corrected Total | 31 | 6425.96875 | | | |

| Root MSE | 14.18082 | R-Square | 0.0612 |
|----------------|-----------|----------|--------|
| Dependent Mean | 144.53125 | Adj R-Sq | 0.0299 |
| Coeff Var | 9.81160 | | |

| Parameter Estimates | | | | | |
|---------------------|----|--------------------|----------------|---------|---------|
| Variable | DF | Parameter Estimate | Standard Error | t Value | Pr > t |
| Intercept | 1 | 140.80000 | 3.66147 | 38.45 | <.0001 |
| smk | 1 | 7.02353 | 5.02350 | 1.40 | 0.1723 |

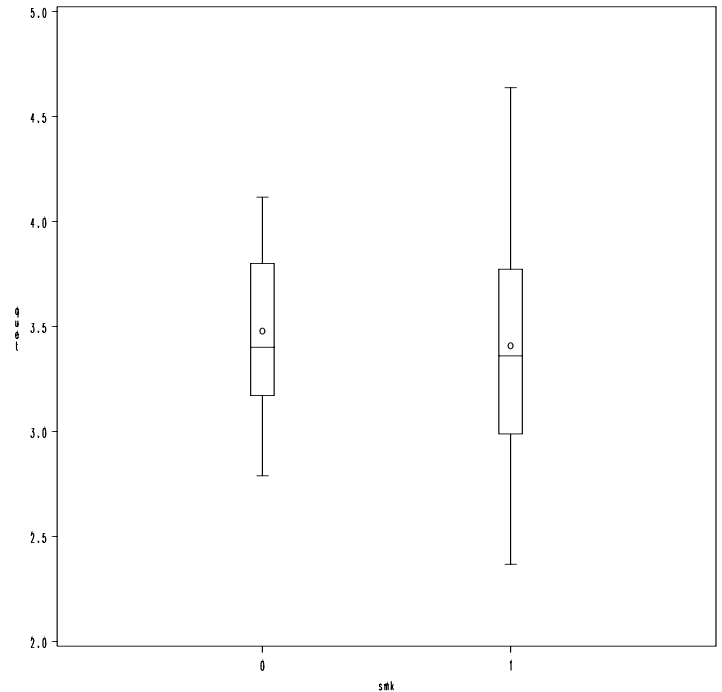
9-13

Boxplots of SBP by SMK



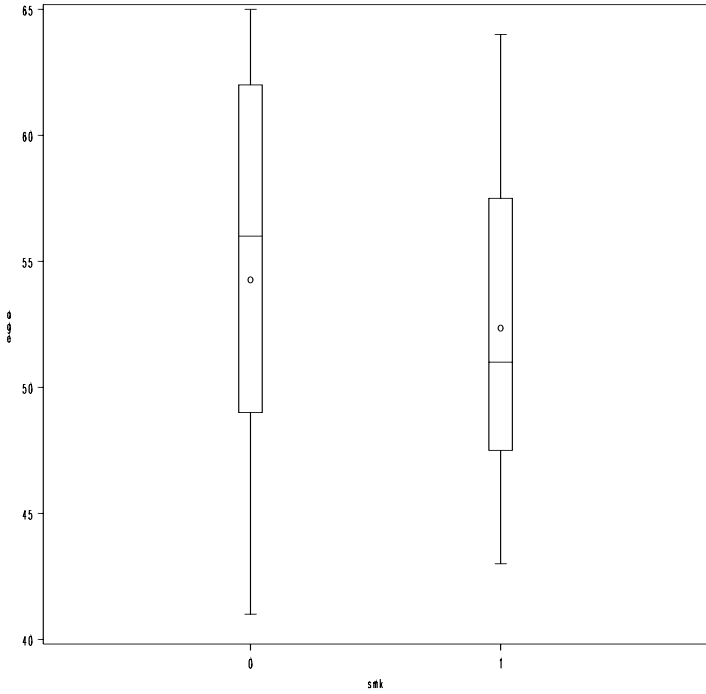
9-14

Boxplots of QUET by SMK



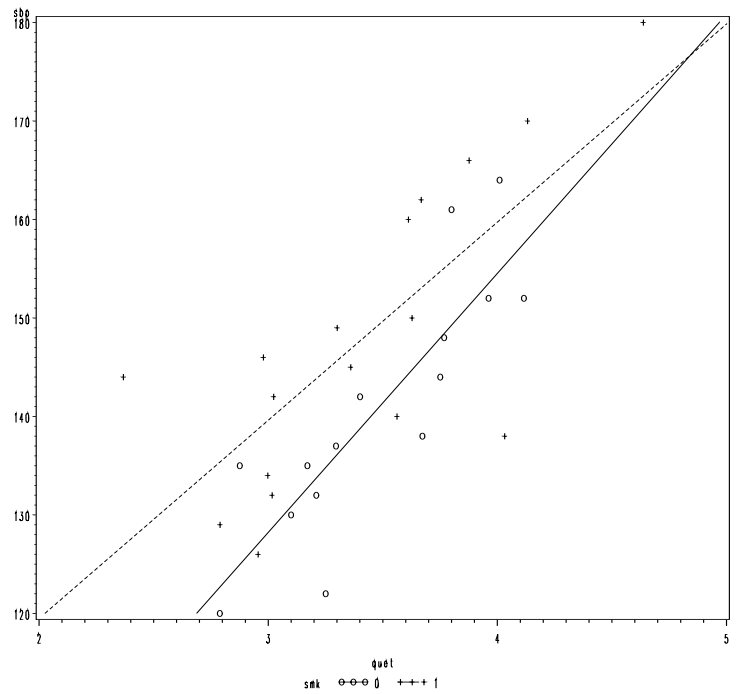
9-15

Boxplots of AGE by SMK



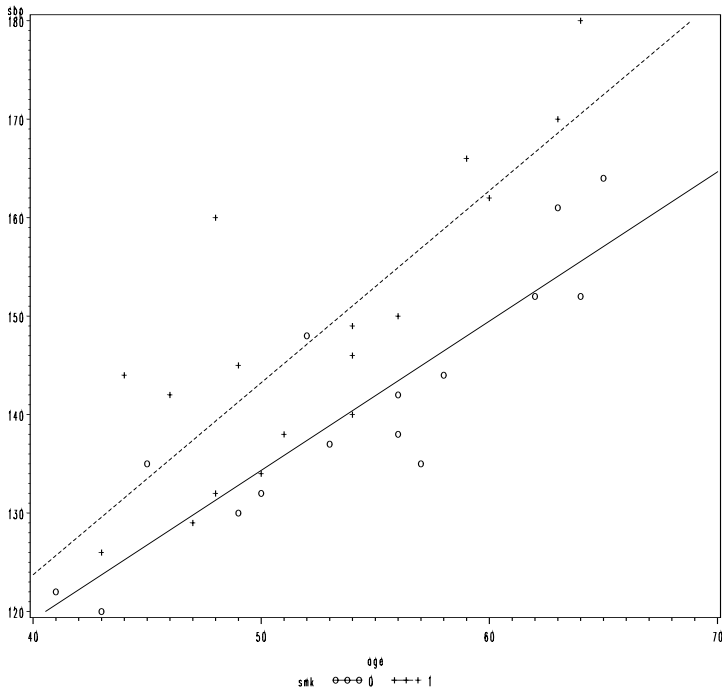
9-16

Scatterplot of SBP vs QUET by SMK



9-17

Scatterplot of SBP vs AGE by SMK



9-18

The REG Procedure

**** Full model

| Analysis of Variance | | | | | |
|----------------------|----|----------------|-------------|---------|--------|
| Source | DF | Sum of Squares | Mean Square | F Value | Pr > F |
| Model | 5 | 5081.86742 | 1016.37348 | 19.66 | <.0001 |
| Error | 26 | 1344.10133 | 51.69621 | | |
| Corrected Total | 31 | 6425.96875 | | | |

| | | | |
|----------------|-----------|----------|--------|
| Root MSE | 7.19001 | R-Square | 0.7908 |
| Dependent Mean | 144.53125 | Adj R-Sq | 0.7506 |
| Coeff Var | 4.97471 | | |

| Parameter Estimates | | | | | |
|---------------------|----|--------------------|----------------|---------|---------|
| Variable | DF | Parameter Estimate | Standard Error | t Value | Pr > t |
| Intercept | 1 | 185.69424 | 73.70952 | 2.52 | 0.0182 |
| quet | 1 | -32.41001 | 21.84299 | -1.48 | 0.1499 |
| age | 1 | -1.35438 | 1.40388 | -0.96 | 0.3436 |
| smk | 1 | 3.46636 | 20.61748 | 0.17 | 0.8678 |
| quet_smk | 1 | 1.80295 | 5.95779 | 0.30 | 0.7646 |
| quet_age | 1 | 0.73888 | 0.38735 | 1.91 | 0.0676 |

The quet_age term in this model implies that the difference between average sbp at two quet values is larger the older the person is.

Notice on the next page how the significance of smk changes when quet_smk terms is removed. This is an example of collinearity.

9-19

Analysis of Variance

| Source | DF | Sum of Squares | Mean Square | F Value | Pr > F |
|-----------------|----|----------------|-------------|---------|--------|
| Model | 4 | 5077.13311 | 1269.28328 | 25.41 | <.0001 |
| Error | 27 | 1348.83564 | 49.95688 | | |
| Corrected Total | 31 | 6425.96875 | | | |

| | | | |
|----------------|-----------|----------|--------|
| Root MSE | 7.06802 | R-Square | 0.7901 |
| Dependent Mean | 144.53125 | Adj R-Sq | 0.7590 |
| Coeff Var | 4.89030 | | |

Parameter Estimates

| Variable | DF | Parameter Estimate | Standard Error | t Value | Pr > t |
|-----------|----|--------------------|----------------|---------|---------|
| Intercept | 1 | 181.70377 | 71.29001 | 2.55 | 0.0168 |
| quet | 1 | -30.69963 | 20.74113 | -1.48 | 0.1504 |
| age | 1 | -1.38470 | 1.37654 | -1.01 | 0.3234 |
| smk | 1 | 9.65649 | 2.53893 | 3.80 | 0.0007 |
| quet_age | 1 | 0.73724 | 0.38074 | 1.94 | 0.0634 |

Keeping everything else fixed, this model says smokers have on average a SBP that is 9.7 units higher than nonsmokers.

The parameter estimates for quet_age, quet, and age have not changed very much.

Common practice to not remove lower order term like quet or age if a higher order term is still in the model.

*** Remove quet_age

Analysis of Variance

| Source | DF | Sum of Squares | Mean Square | F Value | Pr > F |
|-----------------|----|----------------|-------------|---------|--------|
| Model | 3 | 4889.82570 | 1629.94190 | 29.71 | <.0001 |
| Error | 28 | 1536.14305 | 54.86225 | | |
| Corrected Total | 31 | 6425.96875 | | | |

| | | | |
|----------------|-----------|----------|--------|
| Root MSE | 7.40691 | R-Square | 0.7609 |
| Dependent Mean | 144.53125 | Adj R-Sq | 0.7353 |
| Coeff Var | 5.12478 | | |

| Variable | DF | Parameter Estimate | Standard Error | t Value | Pr > t |
|-----------|----|--------------------|----------------|---------|---------|
| Intercept | 1 | 45.10319 | 10.76488 | 4.19 | 0.0003 |
| quet | 1 | 8.59245 | 4.49868 | 1.91 | 0.0664 |
| age | 1 | 1.21271 | 0.32382 | 3.75 | 0.0008 |
| smk | 1 | 9.94557 | 2.65606 | 3.74 | 0.0008 |

***Remove quet

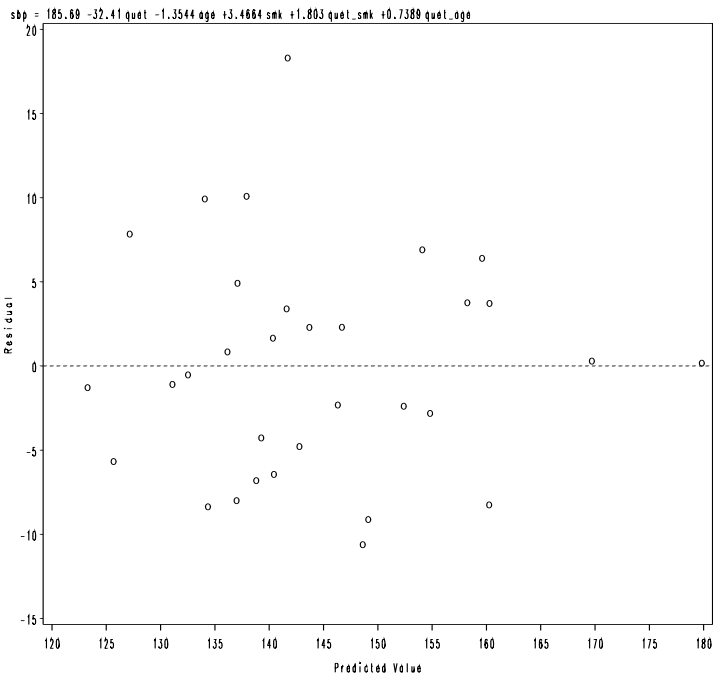
Analysis of Variance

| Source | DF | Sum of Squares | Mean Square | F Value | Pr > F |
|-----------------|----|----------------|-------------|---------|--------|
| Model | 2 | 4689.68423 | 2344.84211 | 39.16 | <.0001 |
| Error | 29 | 1736.28452 | 59.87188 | | |
| Corrected Total | 31 | 6425.96875 | | | |

| | | | |
|----------------|-----------|----------|--------|
| Root MSE | 7.73769 | R-Square | 0.7298 |
| Dependent Mean | 144.53125 | Adj R-Sq | 0.7112 |
| Coeff Var | 5.35365 | | |

| Variable | DF | Parameter Estimate | Standard Error | t Value | Pr > t |
|-----------|----|--------------------|----------------|---------|---------|
| Intercept | 1 | 48.04960 | 11.12956 | 4.32 | 0.0002 |
| age | 1 | 1.70916 | 0.20176 | 8.47 | <.0001 |
| smk | 1 | 10.29439 | 2.76811 | 3.72 | 0.0009 |

Residual plot for full model



Normal QQplot of residuals

